ARMY RESEARCH LABORATORY

The Word-Based Pyramid Method

by Andrew A. Thompson

ARL-TR-5912 **February 2012**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

ARL-TR-5912 **February 2012**

The Word-Based Pyramid Method

Andrew A. Thompson
Weapons and Materials Research Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
February 2012	Final	1 January 2011–9 January 2011

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Word-Based Pyramid Method	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Andrew A. Thompson	AH80
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Research Laboratory ATTN: RDRL-WML-A Aberdeen Proving Ground, MD 21005-5066	ARL-TR-5912

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The comparison of human and machine generated descriptions is a problem confronting the document-understanding community and other researchers developing systems that output natural language. An investigator often wants to know how similar a machine description is to a human description. By assuming the message can be portrayed by a "bag of words," it is possible to automate model formulation and then evaluate descriptions. Based on the study performed, the word-based pyramid method provides a useful tool for the quantification of verbal similarity.

15. SUBJECT TERMS

document understanding, verbal quantification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Andrew A. Thompson
Unclassified	Unclassified	Unclassified	UU	24	19b. TELEPHONE NUMBER (*Include area code*) 410-278-6805

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Acknowledgments

The author would like to recognize Nicholas Fung, Cynthia Pierce, and Laurel Sadler for their exchange of ideas related to aspects of this report. Also, Nicholas Fung's and Laurel Sadler's ability to extract and provide data enabled a timely completion of the task. Cynthia Pierce provided comments that enhanced the quality of this report.

INTENTIONALLY LEFT BLANK.

1. Introduction

The comparison of human and machine generated descriptions is a problem confronting the document-understanding community and other researchers developing systems that output natural language. An investigator often wants to know how similar a machine description is to a human description. In many situations, intelligence reports and situational reports are generated as system output, and it is desired to have a quantitative measure of human similarity. The pyramid method has been used successfully in this domain since 2004. In this method, units of information are defined by human evaluators. These information units or summary content units are then used to form summary models. The process requires human evaluation at all levels and hence is labor intensive and not suitable for large data sets. A potential problem with the pyramid method is the requirement for subjective evaluation of summary content units, possibly leading into discussions of issues associated with objectivity. This report proposes an approach similar to the pyramid method based on words rather than summary content units. By assuming the message can be portrayed by a "bag of words," it is possible to automate model formulation and then evaluate descriptions. While this method is not as precise as the pyramid method, the ability to quickly evaluate a large number of messages will offset this disadvantage in many situations.

The use of this method requires the development of a model derived from human descriptions for each specific document or event. Each model is considered to represent the population from which it is drawn; so for inferences to be made, the issues associated with random sampling are important. For example, if a model for the description of a scene is considered representative of the general group or population, the number of participants contributing to the model increases the generality of the model. Another issue relates to the discrimination of the model. The model should produce low scores for messages or summaries of dissimilar material, and the scores should increase as the messages relate more to the subject matter of the model. It should be noted that these methods seek to measure similarity and not the correctness or truthfulness of the responses or messages. When comparing machine verbalizations with human models, a low similarity score does not imply the machine description is wrong but only that it did not respond like the human model. Also note that the models can be made specific to particular groups; e.g., for video interpretation, the human models could be based on the verbal reports of individuals trained in video surveillance. The method can also be used to compare different groups of humans.

2. Background

The complexity of natural language has presented difficulties to the academic community for decades. By restricting a language or a response to be a first-order predicate calculus, many problems can be avoided and reasoning systems can be developed (*1*). The development of similarity measures requires quantification of some of the features of a natural language passage. In this quantification, some information or detail will be lost; this should be a key selection criterion in selecting a similarity metric. Many similarity methods have been implemented and produced good results for given classes of problems.

Vector space approaches represent descriptions as points in multidimensional space. These points can be based on words or semantic content. There are a variety of terms in the literature that refer to semantic content; these terms include factoid, nugget, summary content unit, and semantic content unit. To form a factoid, etc., an individual must parse a passage and break it into portions of semantic interest. Each factoid can then be represented as a dimension or element of a vector. The word approach uses individual words as dimensions. (A discussion of a vector space model by Dr. E. Garcia can be found at http://www.miislita.com/term-vector /term-vector-3.html [*2*]). Vector space methods are good initial models for investigation as the underlying theory is well defined and accepted by the academic community. Typically, after the vectors have been defined, measures of similarity are proposed; these measures vary across methods. A potential problem for vector space methods exists as the passages get large; this can strain the method as the size of the vectors or dimensionality gets too large for computation or comprehension. Similarity measures tend to have low values in high-dimensional spaces. Features or dimensions of interest for linguistic quality have been discussed by Pitler et al. (*3*).

Ngram approaches are basically Markov chains representing word phrases. Probabilities for successive words or phrases can be calculated and used to develop models. While Ngram approaches are not appealing from the language understanding perspective, they have been useful within the experimental community for developing similarity measures and summary models. The software ROUGE will produce Ngrams for passage analysis (in addition to other metrics). Another software tool that disserves mention is BLEU. BLEU was designed to evaluate the quality of text that has been machine translated. BLEU performs poorly when used to evaluated sentences or short descriptions.

The pyramid method was developed based on semantic content units or summary content units (*4*). This method is based on the occurrences of semantic content units within human summaries. The semantic content unit or summary content unit is similar to the idea of a factoid or nugget used in other methods. These are basic units of information and can typically be thought of as phrases or short sentences. These units can vary based on the subject matter. An instructive example of the pyramid method demonstrates the formation of semantic units and the

development of weights for the model (*5*). To produce a human model for a given document, get "n" humans to summarize the document and then break each summarization into content units. Finally, form a model by giving a weight to each of the factoids or summary content units. The weight will be the number of summaries the factoid is contained within, so weights will be between 1 and n. To evaluate a summary, simply add the weights associated with each factoid contained with the summary. Note that factoids that are not part of the model will be assigned values of zero.

There are articles that review the state of document summary (*6*). The many successful applications of the pyramid method attest to its utility. The current recommendation is that five human summaries give stable results for model generation (*5*).

3. Method

Five individuals from the U.S. Army Research Laboratory (ARL) generated descriptions of 480 videos. The descriptions were typically one or two sentences, with some as short as two words. In order to generate an automated process, words rather than semantic content units were used as the basic unit. A major difference between this data set and data sets associated with articles using the pyramid method is the word length of the individual descriptions. Since 480 models were needed, an automated approach was necessary. The lower word count per message or summary suggested a word-based method could suffice. The pyramid method was applied on words to meet these objectives.

The use of words generates issues that are not consequential when using the pyramid method. As a first step, a dictionary of all the words used was prepared. From this dictionary a list of "stop words" was prepared. Stop words are words that contain limited or no information; they are removed and not used for model generation. The list of stop words generated for this task is not thought to be absolute or unambiguous; it reflects the investigator's perception of stop words related to this set of video descriptions.

Another problem with the use of words relates to synonyms and verb tenses. Different people will select different words with similar meaning to describe events, and although the words are different, the meanings can be the same. This is also true for verb tenses. For example, "the man walks" is the same as "the male walked." Equivalence classes were used to mitigate these problems. Based on the dictionary, 147 equivalence classes were created. The majority of these related to verb tense. The limiting of this task to the dictionary kept this task from getting too cumbersome. The equivalence classes were defined by the investigator. Some problems were left unresolved; for example, is left a direction or the past tense of leave? As an example, consider the two descriptions "the man left" and "the man went left." If "the" and "went" are

considered stop words, the residual comparison is equivalence, even though the meanings are different. The positive aspects of equivalence classes exceed any deficiencies.

Although WordNet was not used in this application, its use was intended. The time to develop the ancillary software to connect with WordNet was estimated to be longer than the time available for the project. The intended use of WordNet, an English language database, is to support automatic text analysis and artificial intelligence applications. VerbNet is a similar tool that focuses on verbs. The use of these applications to create equivalence classes would remove some of the subjectivity associated with the investigator generated equivalence classes.

Each description was parsed, and stop words were removed. Words associated with equivalence classes were then replaced by the equivalence class. Finally, models were generated from the five descriptions associated with each video. The models were generated as indicated by the pyramid method; that is, each word or equivalence class of the model had a weight indicating the number of descriptions it occurred within. The initial results varied significantly; this variability was assumed to be associated with the length of the response. Also, it was noted that verbose responses got higher scores, as a response containing many words had more chances to match the model. As an extreme example, consider if a dictionary was given as the description. If this was the case, the dictionary description would get the highest possible score and be evaluated as a perfect description for any model. It was found that division by the number of retained words in the description resolved these issues (note that this division introduces a penalty for using words that are not in the model). The model evaluations of each description can be thought of as indicating the message agreement per retained word.

4. Investigation of ARL Descriptions

The original data consisted of descriptions of 480 videos by five members of ARL. From these descriptions, 480 word-based pyramid models were automatically generated. Each of the videos was design to exemplify a particular verb from a set of 48. There were different videos as examples of the same verb. Each of the 48 verbs had 10 exemplars or realizations.

The first study was undertaken to examine the sensitivity of the models. The basic question relates to how each model does at identification of similar and dissimilar descriptions or summaries. To determine this, three simulations were completed. In the first, a description was randomly chosen. Then a model was randomly picked for a different verb, and the description was evaluated by that model. In the second simulation, a description was randomly chosen, and a model from the same verb but a different exemplar was randomly selected to evaluate the description. In the third, after a description was randomly chosen, the model for that verb/exemplar was used to evaluate the model. For each simulation, 1000 replications were performed. The range of scores for a description ranged between 0 and 5.

Graphically, the results are presented in figure 1. The green points represent the case where the description was evaluated by a model for a different verb, the dark blue for a similar verb, and the cyan for the same verb/exemplar. It can be seen that while there is not complete separation, there is good separation for the green and cyan points. The models seem to have enough sensitivity to discriminate between similar and dissimilar descriptions.

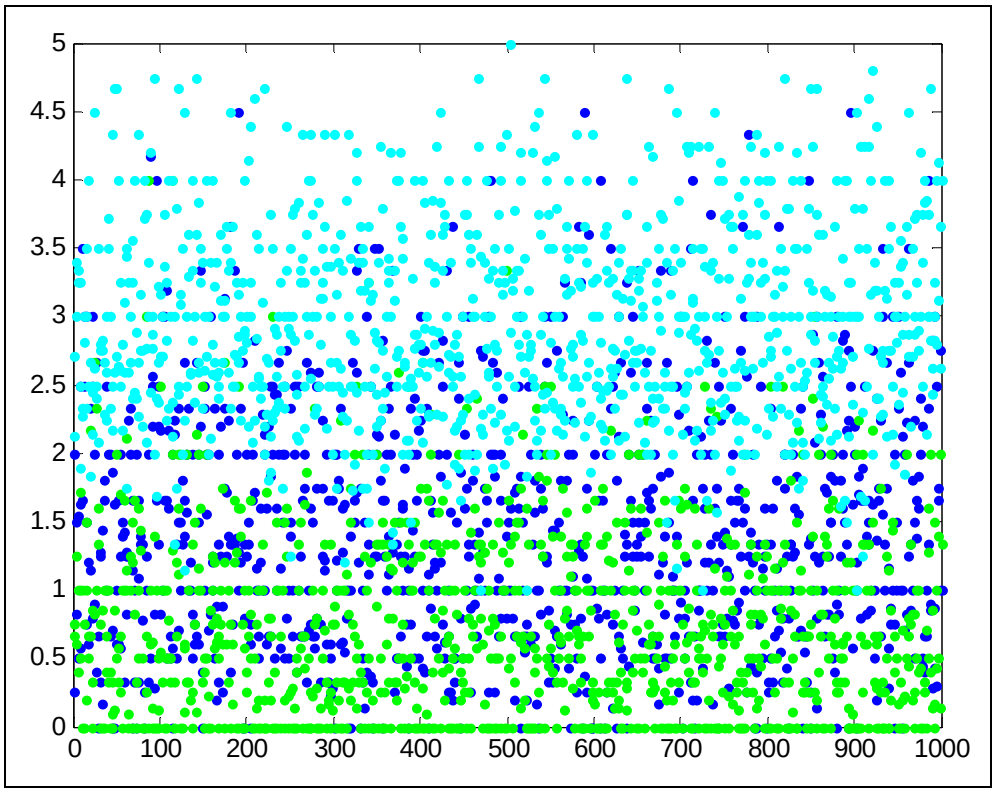

Figure 1. Simulations of descriptions for various models.

If the value 2 is used as the threshold, there are 949 of 1000 $\geqq 2$ of the cyan, so the miss rate is 0.051. For the blue, there are 286 values $\geqq 2$, so the false alarm rate for the blue is 0.286. The green has a false alarm rate of 0.067 when a threshold of 2 is used. From this, it seems the developed method will work. There is still more to do; however, there is currently no evidence against the word-based pyramid method. The word-based pyramid method has reasonable miss and false alarm rates. Both are below 0.07 for a test threshold of 2.

The means and standard deviations based on 1000 samples are given in table 1. The results of this investigation indicate that this can be a good method to evaluate descriptions. The word-based pyramid method can successfully discriminate between descriptions.

Table 1. Means and standard deviations for first simulation.

Color	Mean	Standard Deviation
Cyan	2.96	0.71
Blue	1.43	0.89
Green	0.73	0.67

5. Amazon Turk Study

For each of the 480 previously discussed videos, descriptions were collected from individuals through Amazon Mechanical Turk, one of the suites of Amazon's Web services. Requestors are able to post human intelligence tasks (HITs). Workers can then view the list of HITs, select tasks, and receive monetary compensation. For the study, each HIT was to provide a description of a video. For each of the 480 videos, 15 HITs were acquired. This formed the data set for the Amazon Turk study.

Each description from the Amazon Turk data set was evaluated by the ARL word-based pyramid model for the same verb/exemplar combination. Figure 2 displays these results. The Amazon Turk description evaluations compare similarly with the ARL descriptions for verb/exemplar matches with the model. The mean and standard deviation for the Amazon descriptions are 2.62 and 0.86; these values are close to the ARL results of 2.96 and 0.71 from table 1. The decrease in the mean was expected, as the ARL descriptions were used to generate the ARL models. Also note that the lowest possible value for an ARL description by an ARL model was 1. In light of this information, the decrease in the mean and increase in variance was not unexpected. This change will increase the expected miss rate.

The Amazon Turk data set provides evidence that the ARL models have wide applicability. The miss rate of 0.20 for this data set is about four times larger than that of the ARL data set. In previous studies using Amazon Turk data, it was noted that some workers did not respond in a conscientious manner. If this occurred within this data set, changes to both the mean and variance would occur. Based on the human data from Amazon Turk and ARL, it seems human descriptions are similar as measured by a word-based pyramid model. The high miss rate indicates the sample size needs to be relatively large to attain statistical significance. As evidenced by the 0.2 miss rate, while discrimination is not perfect, it is good. As a result, the models should give good results for reasonable size samples.

For a single case, there is an 80% chance of a correct identification or a hit if the Amazon Turk data represents the general public. If we assume the false alarm rate is similar to the ARL data value of 6.7%, then errors associated with a threshold of 2 will be dominated by the misses rather

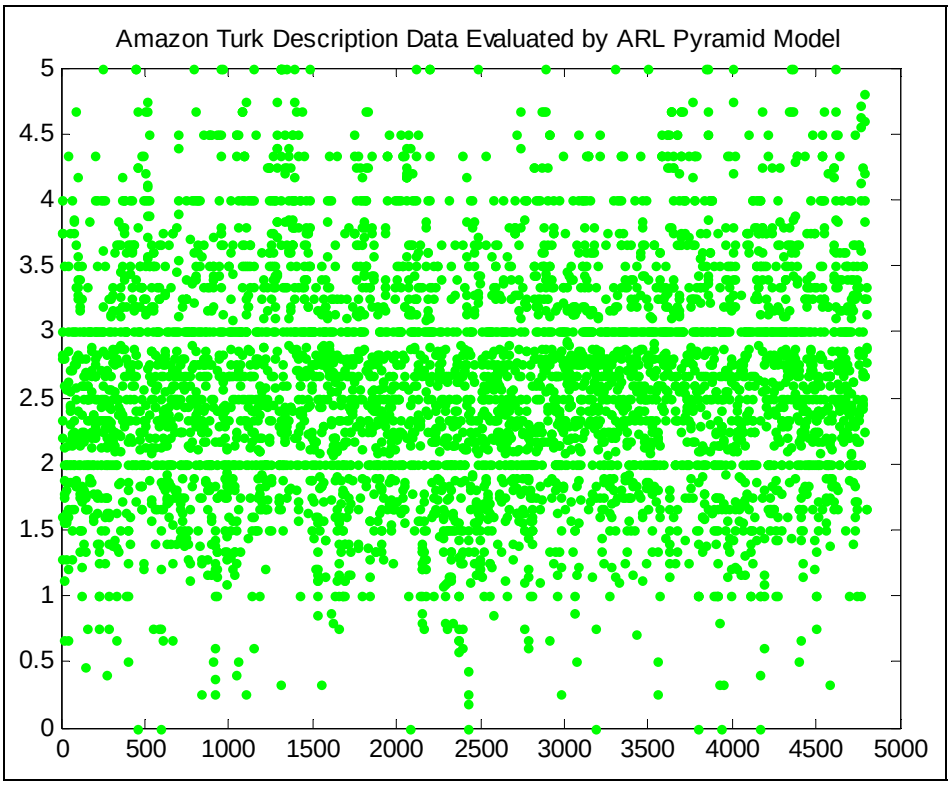

Figure 2. Evaluation of Amazon Turk data.

than the false alarms. From the analysis of human data, it is possible to set up a method that can detect similarity to human performance on a description task; further, this method discriminates between human descriptions of the same video and human descriptions of different videos.

6. Study of Automatic System

An automated video description system provided descriptions for 240 videos. These descriptions were evaluated by the appropriate ARL models. The data is displayed in figure 3. Visually, this data is not similar to the previous sets of data. Both the ARL description data and the Amazon Turk description data have higher means. The mean of the automatic system description data is 1.14, with a standard deviation of 0.81; this is significantly different than either of the other two data sets. Only 42 observations exceed a threshold value of 2 or greater. The automatic system has an extremely high miss rate. An observation of the data from the automatic system indicated that the responses were correct; thus, the choice of language must be different for this system. As evidenced by the low mean and number of description evaluations below the threshold value of 2, the word pyramid method is sensitive enough to discern a difference between system and human video descriptions. In this case, the word choice of the system did not match human word selection.

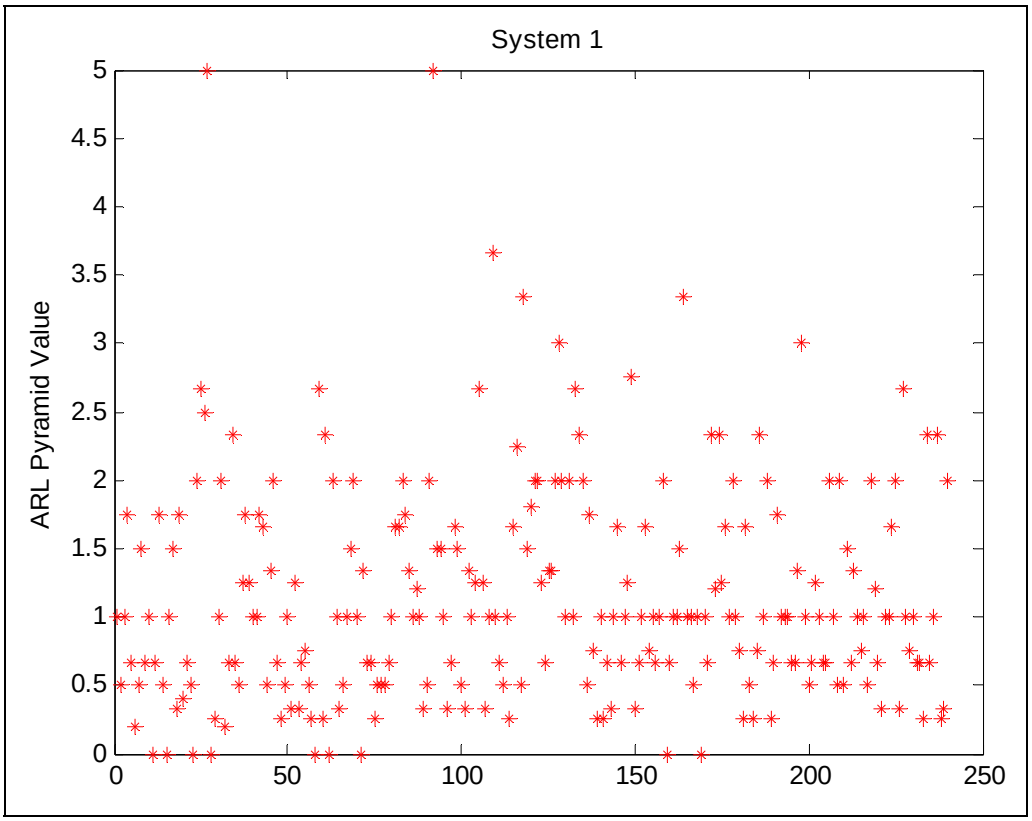

Figure 3. Performance of system 1.

A second system provided 63 responses. The ARL model evaluations are presented in figure 4. The number of description evaluations receiving 0 is high and gives a percentage of 0.48 almost 50%. Mean is 0.902, standard deviation is 1.439, and 11 values are at a threshold of 2 or higher (11/63), or 17% for hit rate. This system did not perform similarly to a human. Observation of the descriptions indicates both the sentence structure and the word choice differ from the human responses. The responses were also not as specific as the human descriptions.

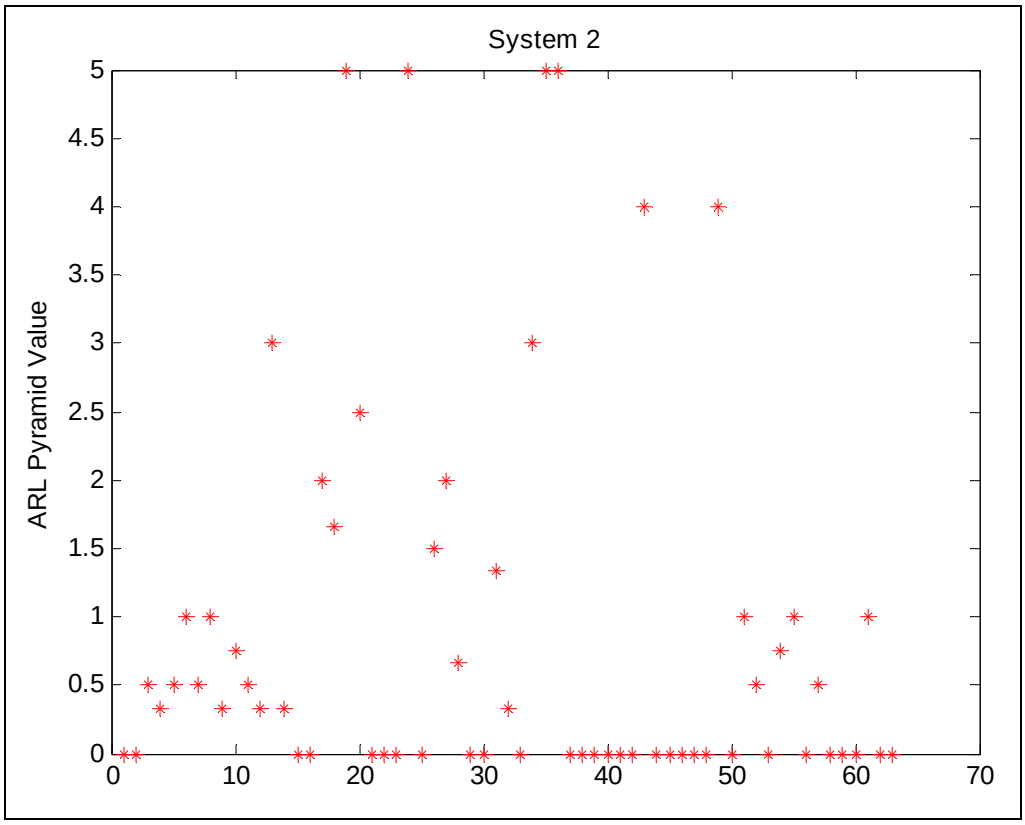

Figure 4. Performance of system 2.

The ARL model evaluations for a third system are presented in figure 5. The percentage of zero descriptions for this system is 11%. The mean is 0.694, with a standard deviation of 0.611. Notice that on this graph, the highest value is 3.5. The hit rate for this system is only 0.0375. Observation of the responses reveals that this system did not identify the objects in the scene as well as humans and the term "object" was contained in almost every response. It also uses a series of short sentences of the standard subject-verb-object form. While this sentence pattern is not similar to the human responses, it is in a form conducive for creating first-order predicate calculus to describe the video, perhaps a precursor for the formation of a video reasoning system.

While all three of these systems provided results of reasonable accuracy, the word-based pyramid test shows that the auto-generated descriptions are not similar to human responses. A major difference is the human description of objects is more specific, that is, the system descriptions tended to use more general verbs and not identify objects within the video. These results demonstrate the difference between similarity, specificity, and accuracy.

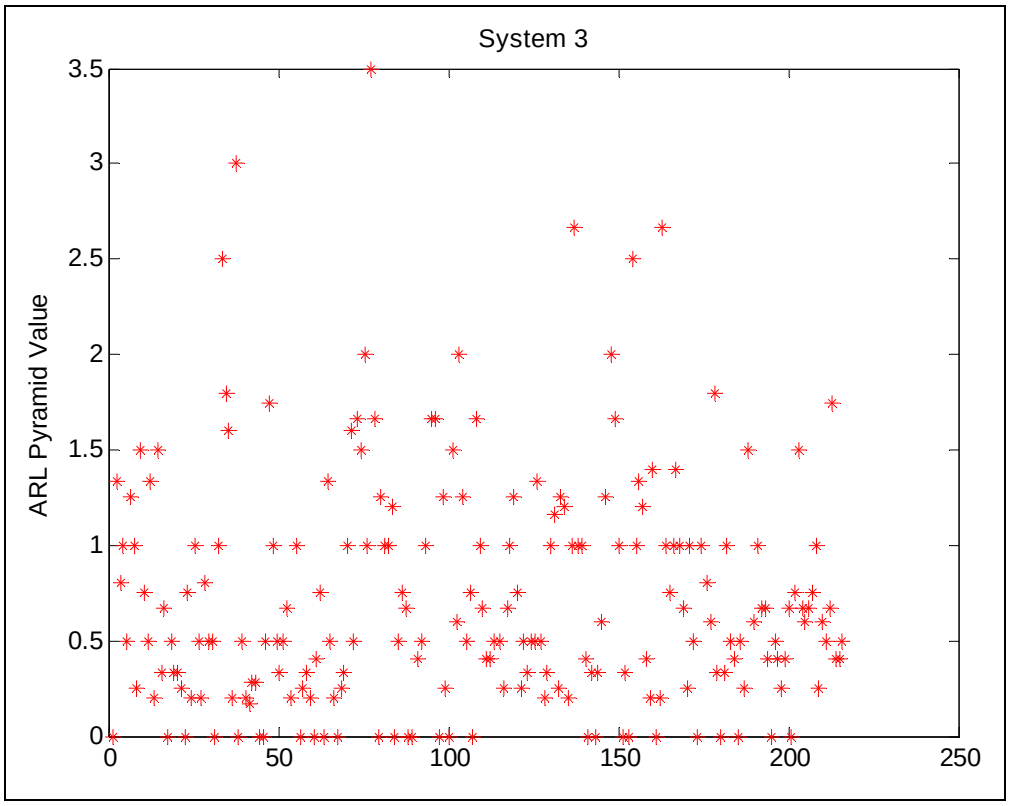

Figure 5. Performance of system 3.

7. Conclusions

The need for the evaluation of verbal descriptions has increased as computing systems translate documents, provide document summaries, and provide verbal output as a system goal. Evaluation of verbal output has been and continues to be a demanding problem. The pyramid method allows the comparison between computer-generated descriptions and human descriptions. This method has been used successfully as the basis of many studies. The document-understanding community typically uses descriptions of several hundred words; in contrast, military descriptions attempt to be terse without loss of information. In many domains, a specialized language exits, and the use of trained personal within a field in conjunction with the word-based pyramid method provides a valuable tool for the evaluation of verbal content from automated systems.

This report investigated a modification of the pyramid method that allowed it to be used automatically. Model generation and description evaluation were both automated. The descriptions evaluated were short by the standards of the document-understanding community, typically less than three sentences. The studies conducted demonstrated that the proposed method can be used successfully to investigate the similarity of verbal responses.

The studies performed provided information indicating that the ARL and Amazon Turk description writers performed in a similar matter. Generalizing slightly, a group of technical experts performed the same as a general group of individuals. This indicates there is similarity in video descriptions across the computer-using community. When performing the same task, an automated system received lower scores for its descriptions. A cursory review indicates the system provides accurate descriptions. This relates to the idea that the pyramid method tests for similarity and not for accuracy.

Besides testing for differences between human and computer generated output, the method can be used to test different human subgroups in their responses to various verbal tasks. Using an automated method typically increases the scope and rate of experimentation in specific domains. Based on the study performed, the word-based pyramid method provides a useful tool for the quantification of verbal similarity.

8. References

1. Davis, E. *Representations of Common Sense Reasoning*; Morgan Kaufmann Publishers: San Francisco, CA, 1990.
2. Garcia, E. The Classic Vector Space Model. http://www.miislita.com/term-vector/term-vector-3.html (accessed October 2006).
3. Pitler, E.; Louis, A.; Nenkova, A. Automatic Evaluation of Linguistic Quality in Multi-document Summarization. Proceedings of 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp 544–554.
4. Nenkova, A.; Passonneau, R. Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics*, Boston, MA, 2–7 May 2004.
5. Nenkova, A.; Passonneau, R.; McKeown, K. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing* **2007**, 4 (*2*), Article 4.
6. Nenkova, A.; McKeown, K. Automatic Summarization. In *Foundations and Trends in Information Retrieval*; 2011; Vol. 5, Nos. 2–3, pp 103–233.

NO. OF COPIES	ORGANIZATION
1 (PDF only)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1	DIRECTOR US ARMY RESEARCH LAB IMNE ALC HRR 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LT 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL D 2800 POWDER MILL RD ADELPHI MD 20783-1197

NO. OF COPIES	ORGANIZATION
1	DEPT OF MECHL ENGRG DREXEL UNIV B C CHANG 3141 CHESTNUT ST PHILADELPHIA PA 19104
3	UNIV OF MARYLAND DEPT OF MECHL ENGRG M MODARRES A MOSLEH J HERRMANN BLDG 088 COLLEGE PARK MD 20742
3	DARPA/I20 J FRITZGERALD P MICHELUCCI J DONLON 3701 N FAIRFAX DR ARLINGTON VA 22203
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIH M D WILSON 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIH N C ADAMS 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LAB RDRL CI D DENT B FORNOFF 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CII B BROOME 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	ORSA CORPORATION J OLAH G THOMPSON 1003 OLD PHILADELPHIA RD ABERDEEN MD 21001
4	DIRECTOR US ARMY RESEARCH LAB RDRL CII A P DAVID N FUNG C PIERCE S YOUNG 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LAB RDRL CII B L SADLER L TOKARCIK 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LAB RDRL CII T C VOSS V HOLLAND 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIN D A CLARK 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIN S C ARNOLD 2800 POWDER MILL RD ADELPHI MD 20783-1197
2	DIRECTOR US ARMY RESEARCH LAB RDRL CIN T R HARDY B RIVERA 2800 POWDER MILL RD ADELPHI MD 20783-1197

NO. OF
COPIES ORGANIZATION

1 UNIV OF PENNSYLVANIA
DEPT OF COMPUTER & INFO SCI
A NENKOVA
3300 WALNUT ST
PHILADELPHIA PA 19104

ABERDEEN PROVING GROUND

1 US ARMY EVALUATION CTR
AEC RAM
A THOMPSON
4120 SUSQUEHANNA AVE
APG

24 DIR USARL
(1 CD) RDRL CIH
R NAMBURU
RDRL CIH M
D WILSON
RDRL CII C
B BODT
J DUMER
A NEIDERER
RDRL CIN D
C ELLIS
L MARVEL
R RESCHLY
RDRL HRS C
E HAAS
RDRL SLB D
J COLLINS
L MOSS
RDRL SLB W
P GILLICH
RDRL WML A
D WEBB
M ARTHUR
A THOMPSON (2 CPS)
B FLANDERS
R PEARSON
B OBERLE (CD ONLY)
RDRL WML F
T HARKINS
M ILG
R MCGEE
RDRL WML H
T BROWN

INTENTIONALLY LEFT BLANK.